

# Breaking the Sample Size Barrier in Model-Based Reinforcement Learning



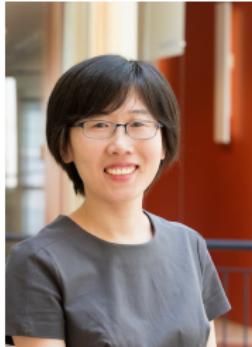
Yuting Wei

Carnegie Mellon University

Nov, 2020



Gen Li  
Tsinghua EE



Yuejie Chi  
CMU ECE



Yuantao Gu  
Tsinghua EE



Yuxin Chen  
Princeton EE

# Reinforcement learning (RL)

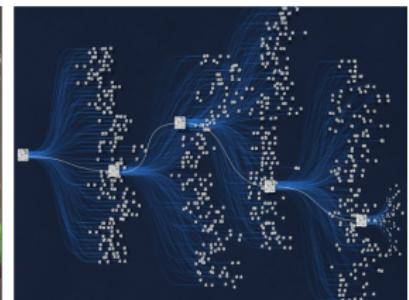
---



# RL challenges

---

- Unknown or changing environment
- Credit assignment problem
- Enormous state and action space



# Provable efficiency

---



- Collecting samples might be expensive or impossible:  
**sample efficiency**
- Training deep RL algorithms might take long time:  
**computational efficiency**

# This talk

---

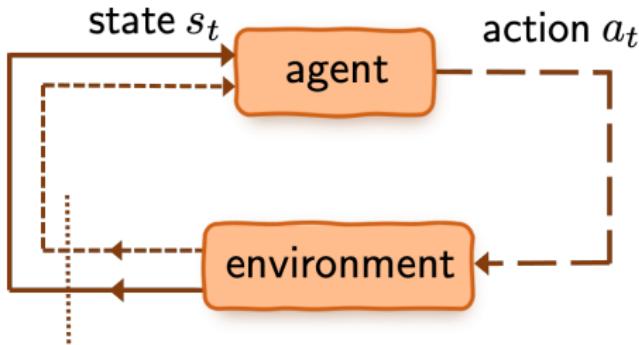
**Question:** can we design **sample- and computation-efficient**  
**RL algorithms?**

— *inspired by numerous prior work*  
[[Kearns and Singh, 1999](#), [Sidford et al., 2018a](#), [Agarwal et al., 2019](#)]...

## **Background: Markov decision processes**

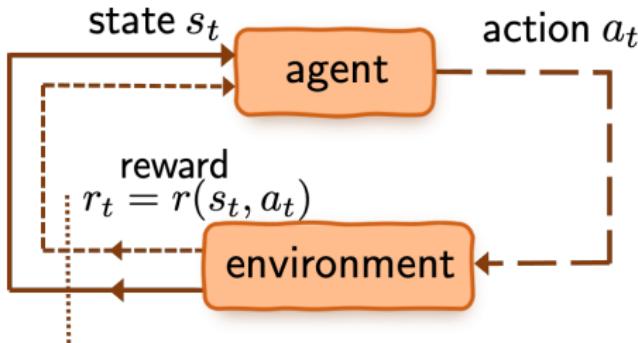
# Markov decision process (MDP)

---



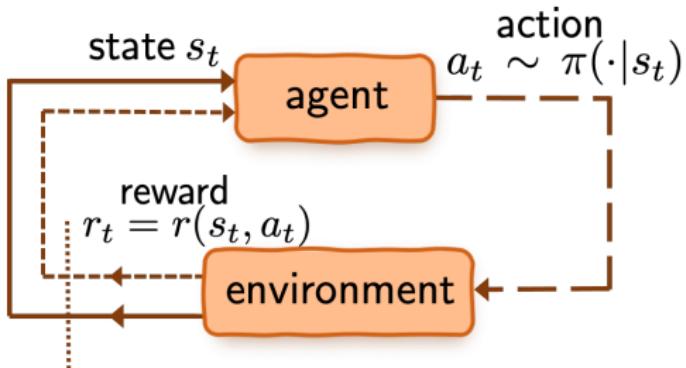
- $\mathcal{S}$ : state space
- $\mathcal{A}$ : action space

# Markov decision process (MDP)



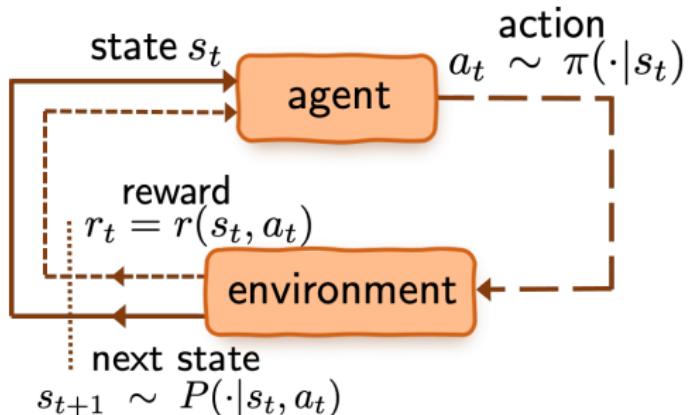
- $\mathcal{S}$ : state space
- $\mathcal{A}$ : action space
- $r(s, a) \in [0, 1]$ : immediate reward

# Markov decision process (MDP)



- $\mathcal{S}$ : state space
- $\mathcal{A}$ : action space
- $r(s, a) \in [0, 1]$ : immediate reward
- $\pi(\cdot | s)$ : policy (or action selection rule)

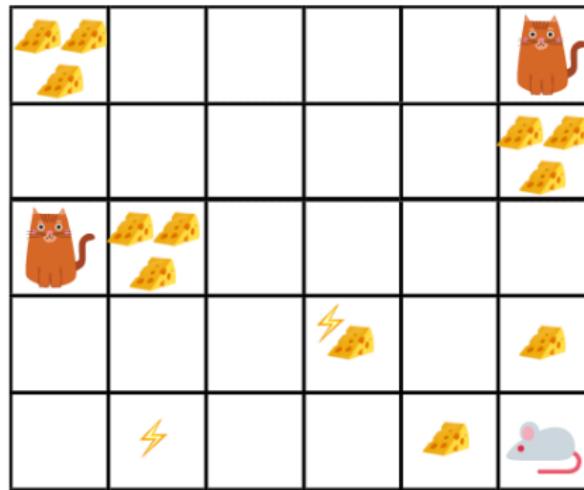
# Markov decision process (MDP)



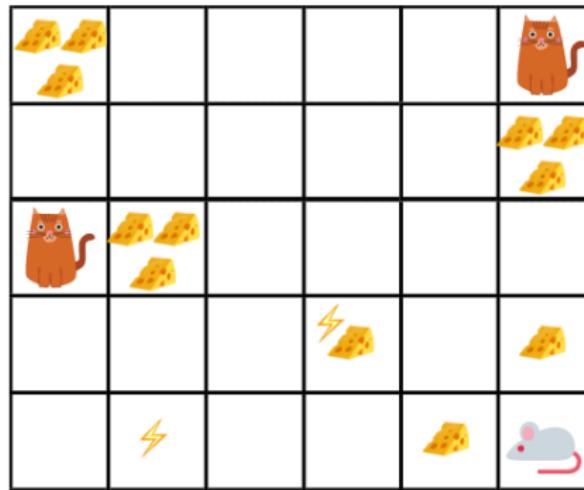
- $\mathcal{S}$ : state space
- $\mathcal{A}$ : action space
- $r(s, a) \in [0, 1]$ : immediate reward
- $\pi(\cdot | s)$ : policy (or action selection rule)
- $P(\cdot | s, a)$ : **unknown** transition probabilities

# Help the rat!

---

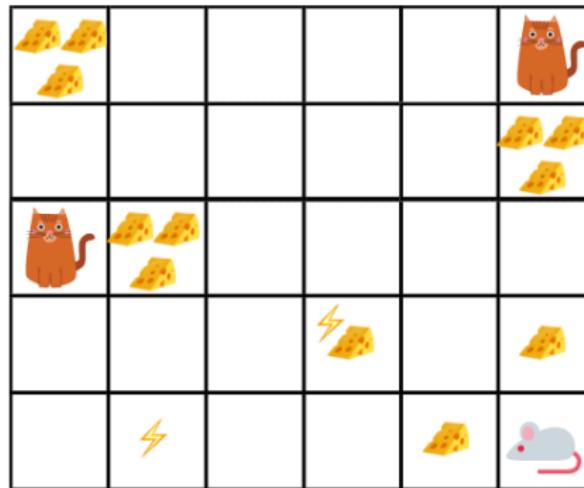


# Help the rat!



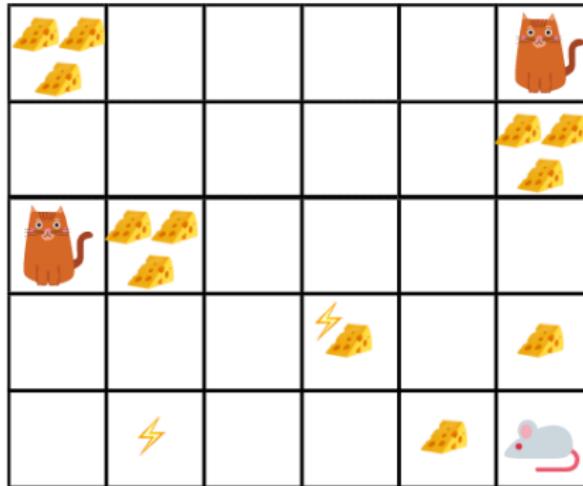
- state space  $\mathcal{S}$ : positions in the maze

# Help the rat!



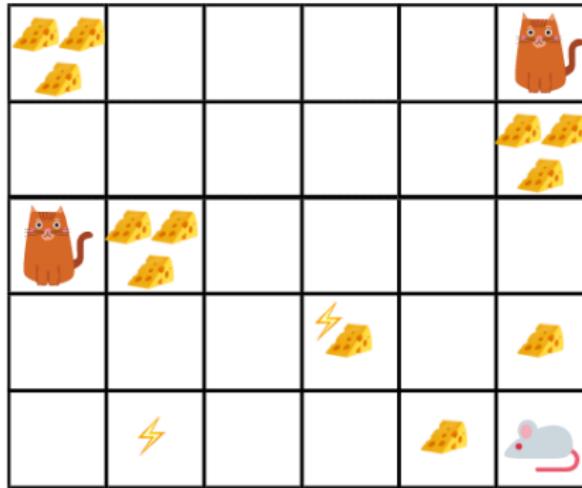
- state space  $\mathcal{S}$ : positions in the maze
- action space  $\mathcal{A}$ : up, down, left, right

# Help the rat!



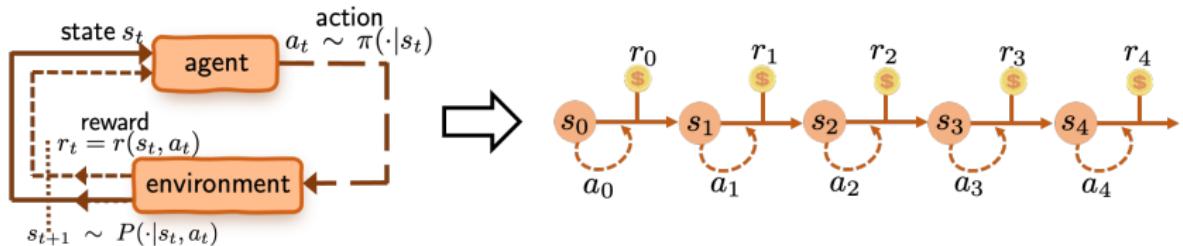
- state space  $\mathcal{S}$ : positions in the maze
- action space  $\mathcal{A}$ : up, down, left, right
- immediate reward  $r$ : cheese, electricity shocks, cats

# Help the rat!



- state space  $\mathcal{S}$ : positions in the maze
- action space  $\mathcal{A}$ : up, down, left, right
- immediate reward  $r$ : cheese, electricity shocks, cats
- policy  $\pi(\cdot|s)$ : the way to find cheese

# Value function

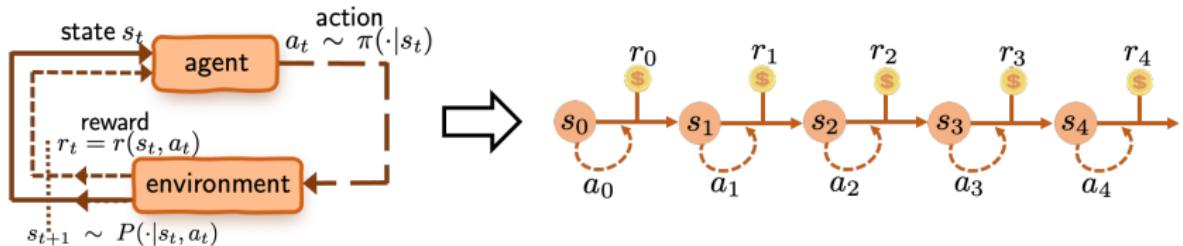


Value function of policy  $\pi$ : long-term **discounted** reward

$$\forall s \in \mathcal{S} : V^\pi(s) := \mathbb{E} \left[ \sum_{t=0}^{\infty} \gamma^t r_t \mid s_0 = s \right]$$



# Value function



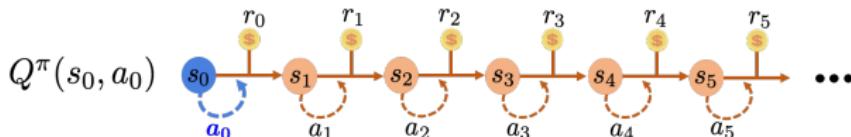
Value function of policy  $\pi$ : long-term **discounted** reward

$$\forall s \in \mathcal{S} : V^\pi(s) := \mathbb{E} \left[ \sum_{t=0}^{\infty} \gamma^t r_t \mid s_0 = s \right]$$



- $\gamma \in [0, 1)$ : discount factor
- $(a_0, s_1, a_1, s_2, a_2, \dots)$ : generated under policy  $\pi$

# Action-value function (a.k.a. Q-function)

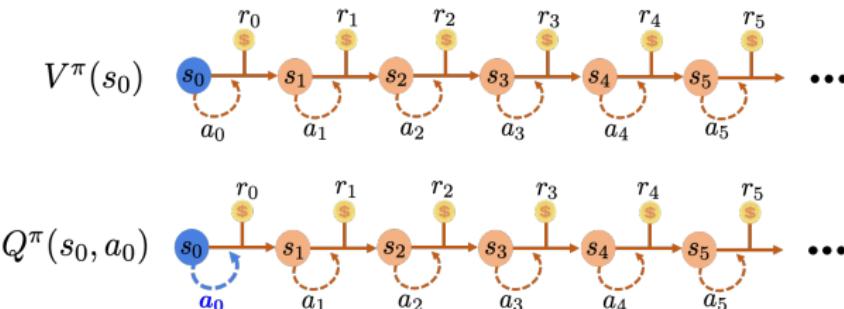


Q-function of policy  $\pi$

$$\forall (s, a) \in \mathcal{S} \times \mathcal{A} : \quad Q^\pi(s, a) := \mathbb{E} \left[ \sum_{t=0}^{\infty} \gamma^t r_t \mid s_0 = s, a_0 = a \right]$$

- ~~$(a_0, s_1, a_1, s_2, a_2, \dots)$~~ : generated under policy  $\pi$

# Action-value function (a.k.a. Q-function)



Q-function of policy  $\pi$

$$\forall (s, a) \in \mathcal{S} \times \mathcal{A} : \quad Q^\pi(s, a) := \mathbb{E} \left[ \sum_{t=0}^{\infty} \gamma^t r_t \mid s_0 = s, a_0 = a \right]$$

- $(\cancel{a_0}, s_1, a_1, s_2, a_2, \dots)$ : generated under policy  $\pi$

# Optimal policy

---



# Optimal policy

---



- **optimal policy**  $\pi^*$ : maximizing value function

# Optimal policy

---

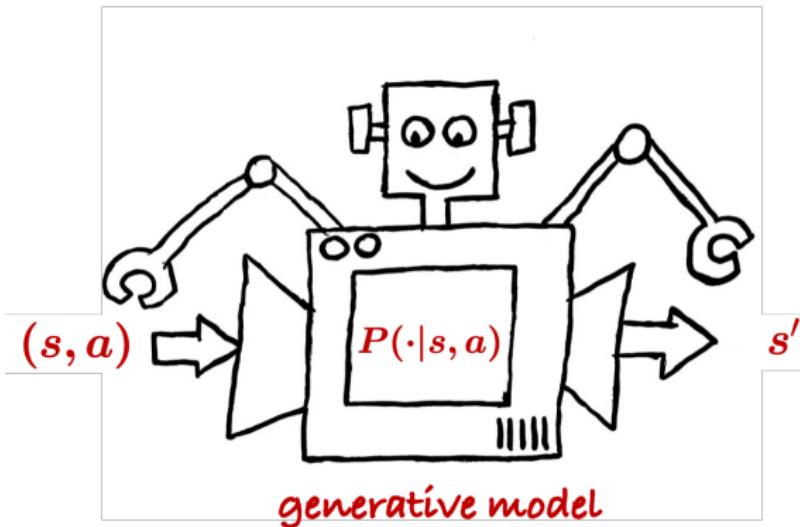


- **optimal policy**  $\pi^*$ : maximizing value function
- optimal value / Q function:  $V^* := V^{\pi^*}$ ;  $Q^* := Q^{\pi^*}$

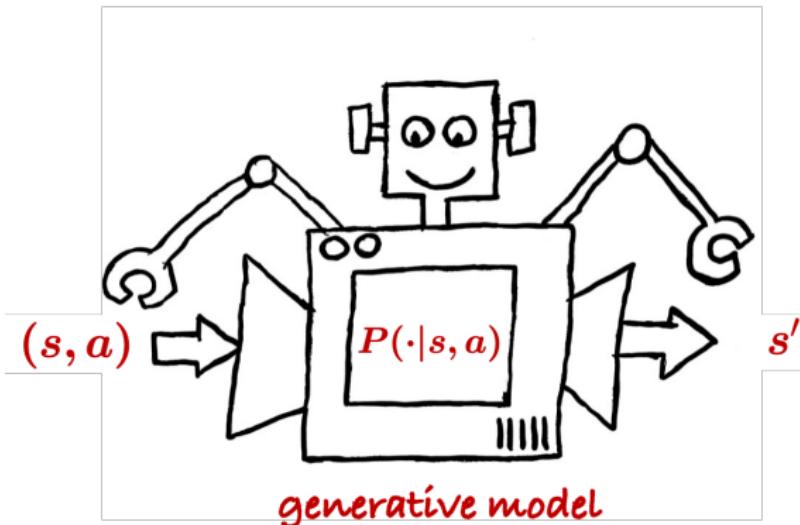
**Practically, learn the optimal policy from data samples ...**

## This talk: sampling from a generative model

---

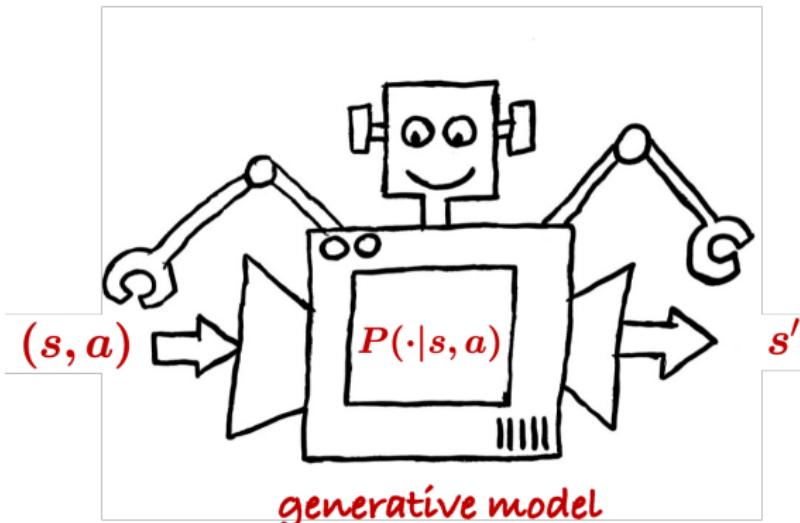


## This talk: sampling from a generative model



For each state-action pair  $(s, a)$ , collect  $N$  samples  $\{(s, a, s'_{(i)})\}_{1 \leq i \leq N}$

## This talk: sampling from a generative model



For each state-action pair  $(s, a)$ , collect  $N$  samples  $\{(s, a, s'_{(i)})\}_{1 \leq i \leq N}$

How many samples are sufficient to learn an  $\varepsilon$ -optimal policy?

## An incomplete list of prior art

---

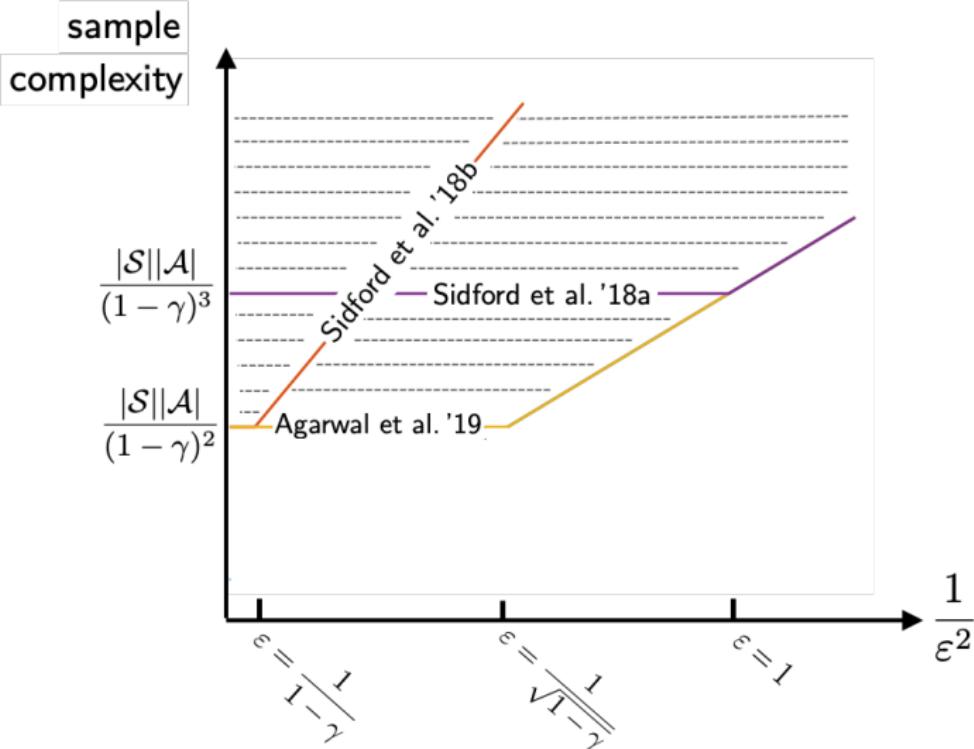
- [Kearns and Singh, 1999]
- [Kakade, 2003]
- [Kearns et al., 2002]
- [Azar et al., 2012]
- [Azar et al., 2013]
- [Sidford et al., 2018a]
- [Sidford et al., 2018b]
- [Wang, 2019]
- [Agarwal et al., 2019]
- [Wainwright, 2019a]
- [Wainwright, 2019b]
- [Pananjady and Wainwright, 2019]
- [Yang and Wang, 2019]
- [Khamaru et al., 2020]
- [Mou et al., 2020]
- ...

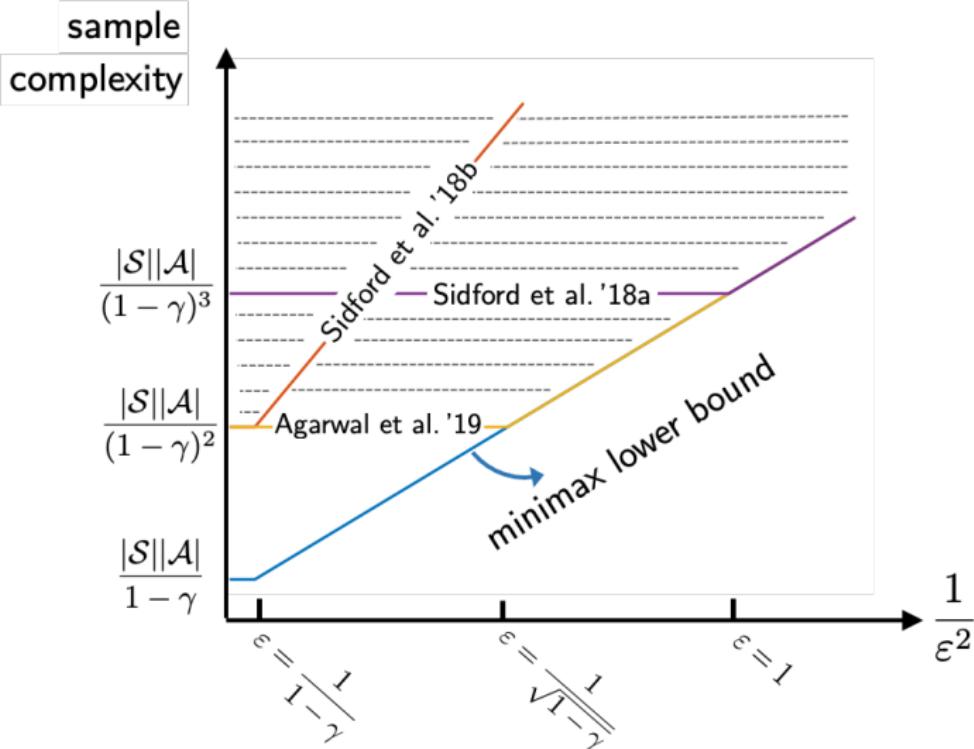
## An even shorter list of prior art

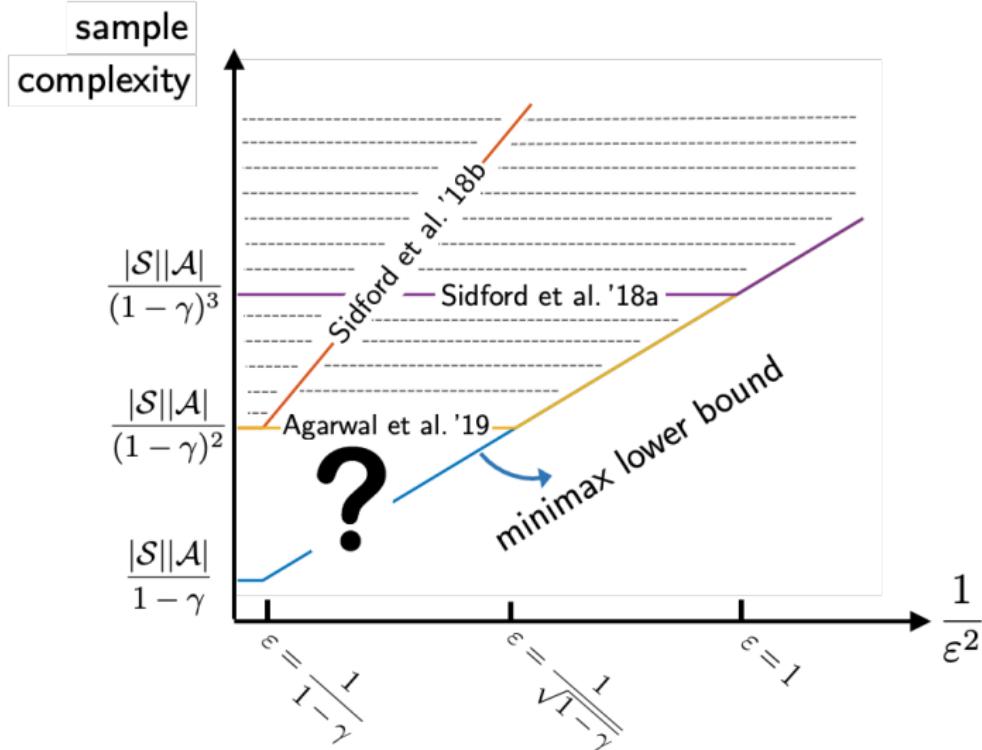
algorithm	sample size range	sample complexity	$\varepsilon$ -range
Empirical QVI [Azar et al., 2013]	$[\frac{ \mathcal{S}  \mathcal{A} }{(1-\gamma)^2}, \infty)$	$\frac{ \mathcal{S}  \mathcal{A} }{(1-\gamma)^3\varepsilon^2}$	$(0, \frac{1}{\sqrt{(1-\gamma) \mathcal{S} }}]$
Sublinear randomized VI [Sidford et al., 2018b]	$[\frac{ \mathcal{S}  \mathcal{A} }{(1-\gamma)^2}, \infty)$	$\frac{ \mathcal{S}  \mathcal{A} }{(1-\gamma)^4\varepsilon^2}$	$(0, \frac{1}{1-\gamma}]$
Variance-reduced QVI [Sidford et al., 2018a]	$[\frac{ \mathcal{S}  \mathcal{A} }{(1-\gamma)^3}, \infty)$	$\frac{ \mathcal{S}  \mathcal{A} }{(1-\gamma)^3\varepsilon^2}$	$(0, 1]$
Randomized primal-dual [Wang, 2019]	$[\frac{ \mathcal{S}  \mathcal{A} }{(1-\gamma)^2}, \infty)$	$\frac{ \mathcal{S}  \mathcal{A} }{(1-\gamma)^4\varepsilon^2}$	$(0, \frac{1}{1-\gamma}]$
Empirical MDP + planning [Agarwal et al., 2019]	$[\frac{ \mathcal{S}  \mathcal{A} }{(1-\gamma)^2}, \infty)$	$\frac{ \mathcal{S}  \mathcal{A} }{(1-\gamma)^3\varepsilon^2}$	$(0, \frac{1}{\sqrt{1-\gamma}}]$

important parameters  $\implies$

- # states  $|\mathcal{S}|$ , # actions  $|\mathcal{A}|$
- the discounted complexity  $\frac{1}{1-\gamma}$
- approximation error  $\varepsilon \in (0, \frac{1}{1-\gamma}]$



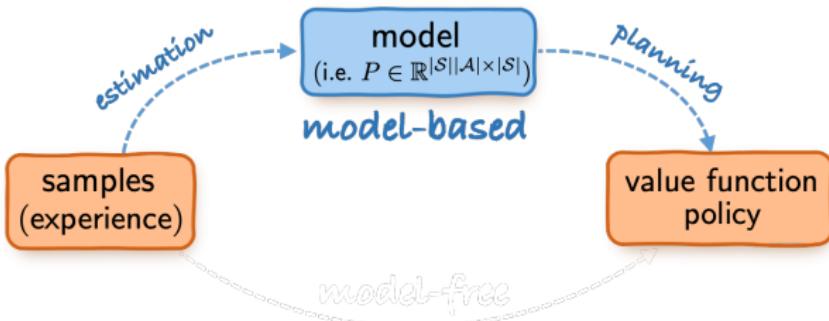




All prior theory requires sample size  $> \underbrace{\frac{|S||\mathcal{A}|}{(1-\gamma)^2}}_{\text{sample size barrier}}$

**This talk: break the sample complexity barrier**

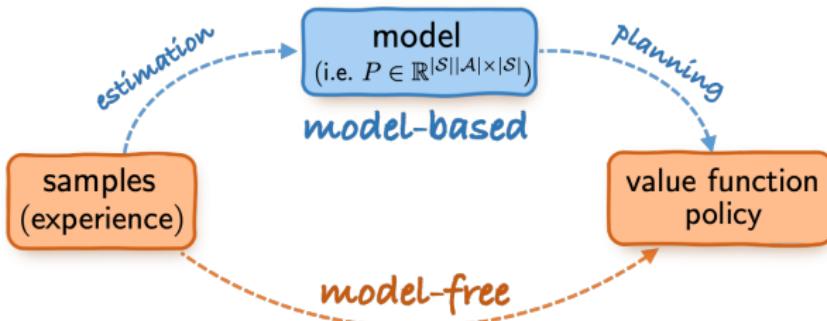
# Two approaches



## Model-based approach (“plug-in”)

1. build empirical estimate  $\hat{P}$  for  $P$
2. planning based on empirical  $\hat{P}$

# Two approaches



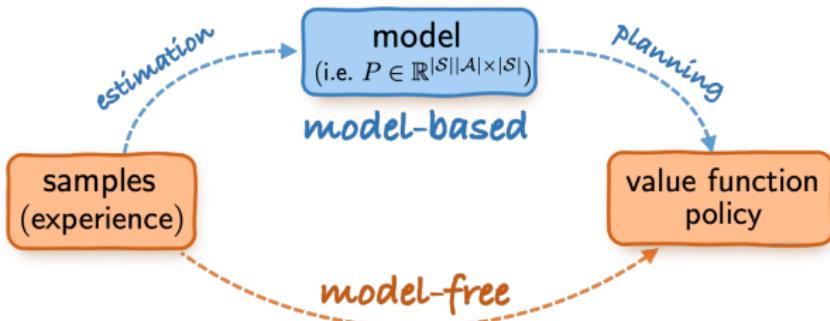
## Model-based approach (“plug-in”)

1. build empirical estimate  $\hat{P}$  for  $P$
2. planning based on empirical  $\hat{P}$

## Model-free approach

— learning w/o constructing a model explicitly

# Two approaches



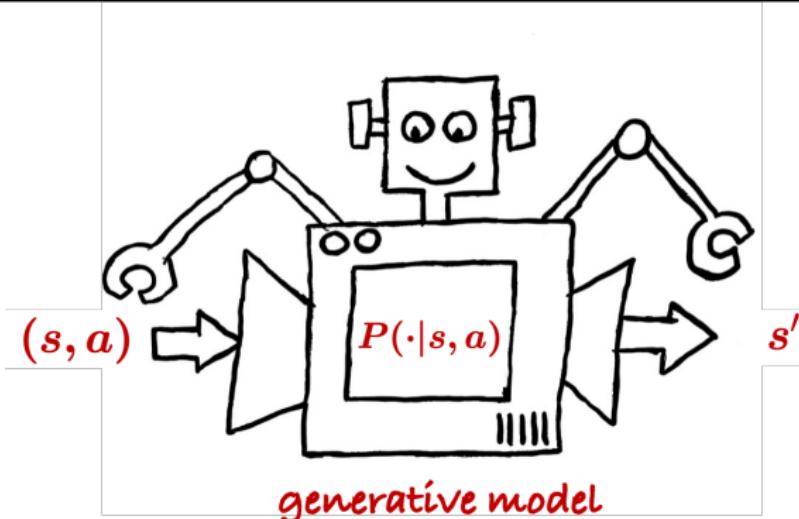
## Model-based approach (“plug-in”)

1. build empirical estimate  $\hat{P}$  for  $P$
2. planning based on empirical  $\hat{P}$

## Model-free approach

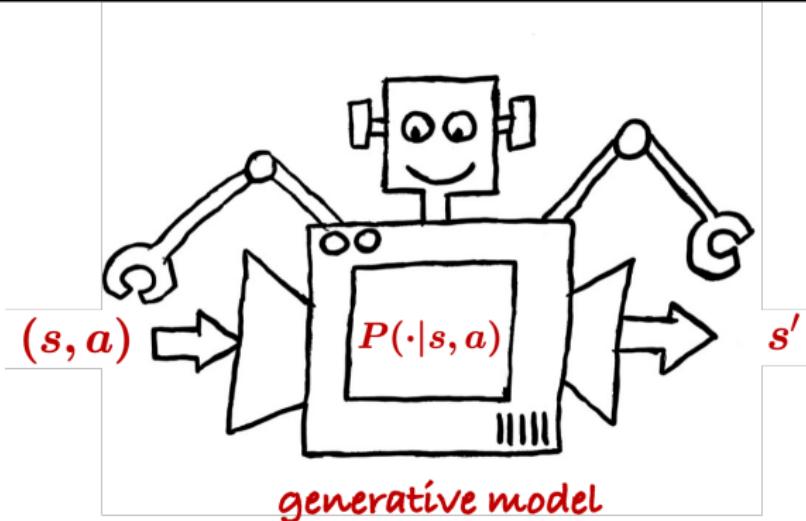
— learning w/o constructing a model explicitly

# Model estimation



**Sampling:** for each  $(s, a)$ , collect  $N$  ind. samples  $\{(s, a, s'_{(i)})\}_{1 \leq i \leq N}$

# Model estimation

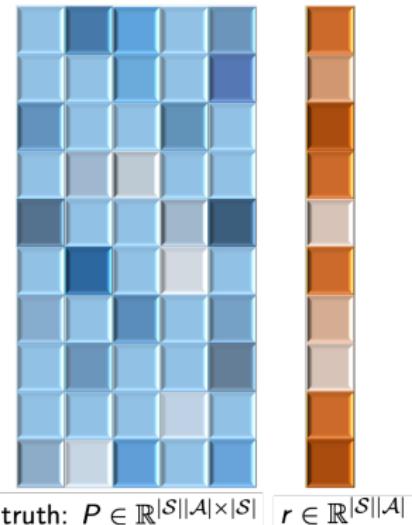


**Sampling:** for each  $(s, a)$ , collect  $N$  ind. samples  $\{(s, a, s'_{(i)})\}_{1 \leq i \leq N}$

**Empirical estimates:** estimate  $\hat{P}(s'|s, a)$  by  $\underbrace{\frac{1}{N} \sum_{i=1}^N \mathbb{1}\{s'_{(i)} = s'\}}_{\text{empirical frequency}}$

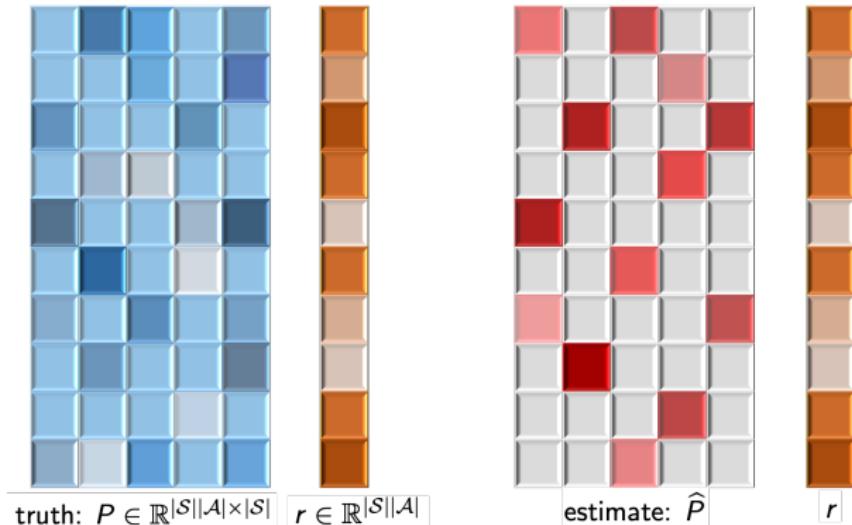
## Our method: plug-in estimator + perturbation

---



original MDP  $(\mathcal{S}, \mathcal{A}, \textcolor{red}{P}, r, \gamma)$        $\iff$        $\pi^* = \arg \max_{\pi} V^{\pi}$

## Our method: plug-in estimator + perturbation



original MDP  $(\mathcal{S}, \mathcal{A}, P, r, \gamma)$



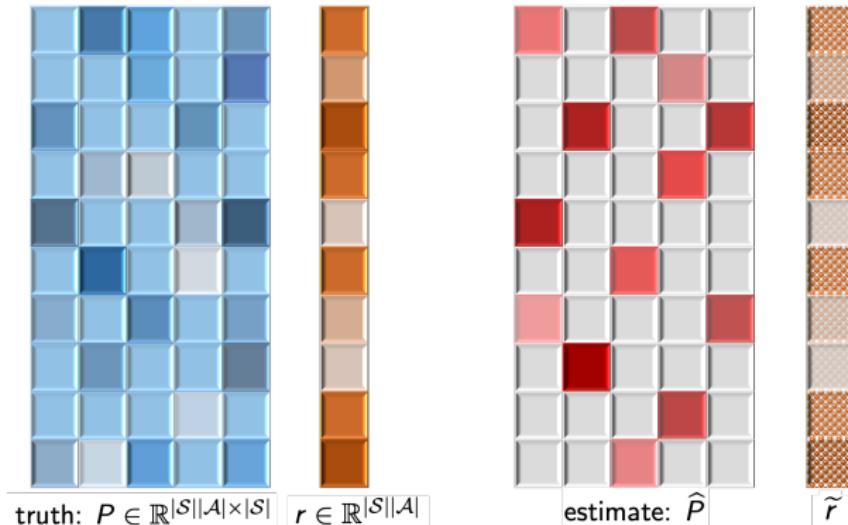
$$\pi^* = \arg \max_{\pi} V^\pi$$

empirical MDP  $(\mathcal{S}, \mathcal{A}, \hat{P}, r, \gamma)$



$$\hat{\pi}^* = \arg \max_{\pi} \hat{V}^\pi$$

## Our method: plug-in estimator + perturbation

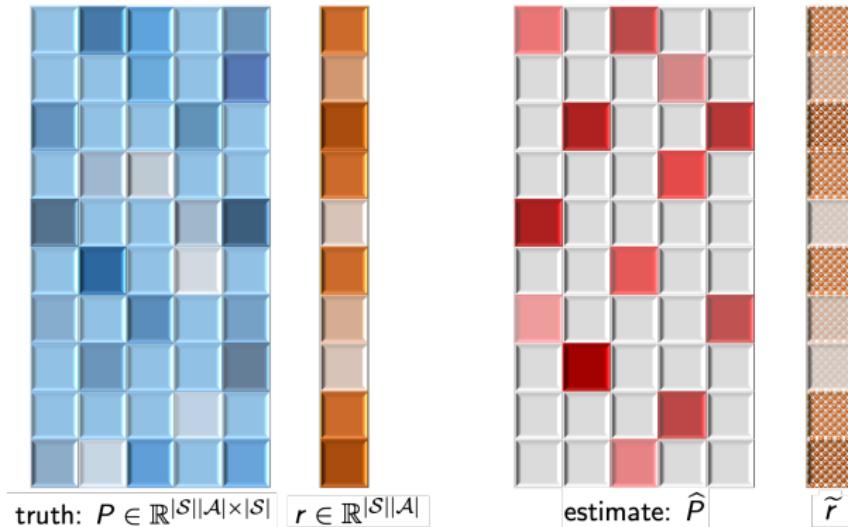


$$\text{original MDP } (\mathcal{S}, \mathcal{A}, \textcolor{red}{P}, r, \gamma) \quad \iff \quad \pi^* = \arg \max_{\pi} V^{\pi}$$

$$\text{empirical MDP } (\mathcal{S}, \mathcal{A}, \widehat{P}, r, \gamma) \quad \iff \quad \widehat{\pi}^* = \arg \max_{\pi} \widehat{V}^{\pi}$$

$$\text{perturbed MDP } (\mathcal{S}, \mathcal{A}, \widehat{P}, \tilde{r}, \gamma) \quad \iff \quad \widehat{\pi}_{\text{p}}^* = \arg \max_{\pi} \widehat{V}_{\text{p}}^{\pi}$$

# Our method: plug-in estimator + perturbation



$$\text{original MDP } (S, \mathcal{A}, P, r, \gamma) \quad \Rightarrow \quad \pi^* = \arg \max_{\pi} V^{\pi}$$

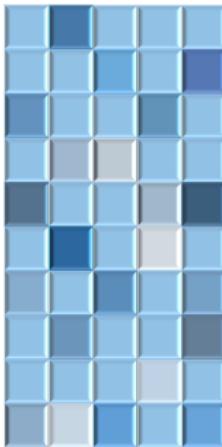
$$\text{empirical MDP } (S, \mathcal{A}, \hat{P}, r, \gamma) \quad \Rightarrow \quad \hat{\pi}^* = \arg \max_{\pi} \hat{V}^{\pi}$$

$$\text{perturbed MDP } (S, \mathcal{A}, \hat{P}, \tilde{r}, \gamma) \quad \Rightarrow \quad \underbrace{\pi_p^* = \arg \max_{\pi} \hat{V}_p^{\pi}}$$

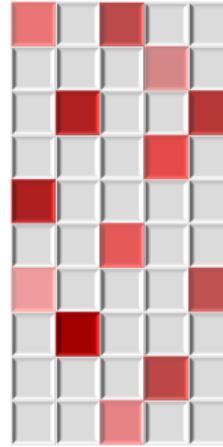
planning (policy iteration, Q-value iteration, ...)

## Challenges in the sample-starved regime

---



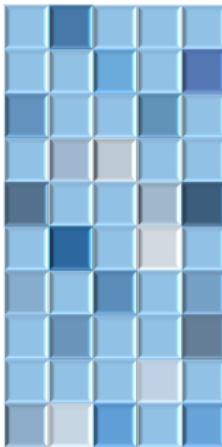
truth:  $P \in \mathbb{R}^{|S||\mathcal{A}| \times |S|}$



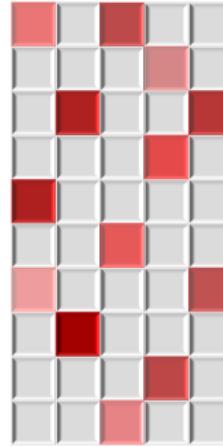
empirical estimate:  $\hat{P}$

- can't recover  $P$  faithfully if sample size  $\ll |S|^2|\mathcal{A}|!$

## Challenges in the sample-starved regime



truth:  $P \in \mathbb{R}^{|S||\mathcal{A}| \times |S|}$



empirical estimate:  $\hat{P}$

- can't recover  $P$  faithfully if sample size  $\ll |S|^2|\mathcal{A}|!$

Can we trust our policy estimate when reliable model estimation is infeasible?

## Main result

---

### Theorem (Li, Wei, Chi, Gu, Chen '20)

For every  $0 < \varepsilon \leq \frac{1}{1-\gamma}$ , policy  $\widehat{\pi}_p^*$  of perturbed empirical MDP achieves

$$\|V^{\widehat{\pi}_p^*} - V^*\|_\infty \leq \varepsilon \quad \text{and} \quad \|Q^{\widehat{\pi}_p^*} - Q^*\|_\infty \leq \gamma\varepsilon$$

with sample complexity at most

$$\tilde{O}\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\varepsilon^2}\right).$$

## Main result

### Theorem (Li, Wei, Chi, Gu, Chen '20)

For every  $0 < \varepsilon \leq \frac{1}{1-\gamma}$ , policy  $\widehat{\pi}_p^*$  of perturbed empirical MDP achieves

$$\|V^{\widehat{\pi}_p^*} - V^*\|_\infty \leq \varepsilon \quad \text{and} \quad \|Q^{\widehat{\pi}_p^*} - Q^*\|_\infty \leq \gamma\varepsilon$$

with sample complexity at most

$$\tilde{O}\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\varepsilon^2}\right).$$

- $\widehat{\pi}_p^*$ : obtained by empirical QVI or PI within  $\tilde{O}\left(\frac{1}{1-\gamma}\right)$  iterations

## Main result

### Theorem (Li, Wei, Chi, Gu, Chen '20)

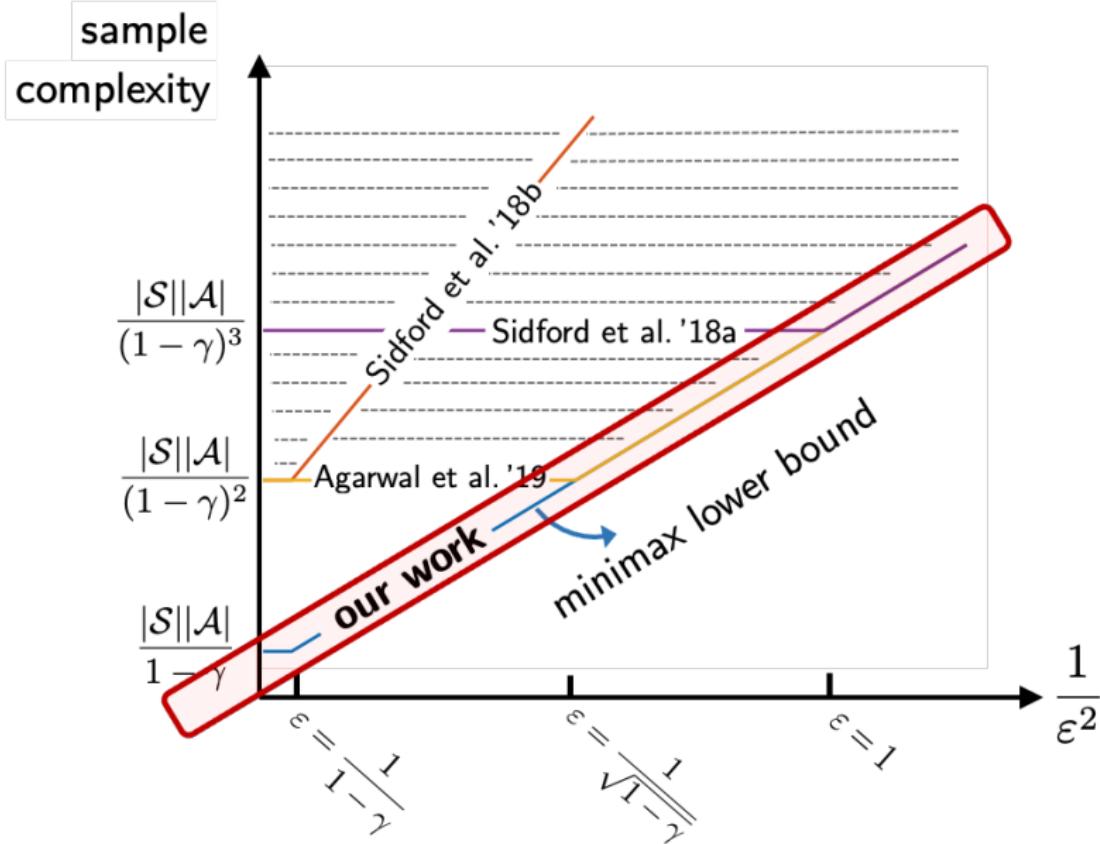
For every  $0 < \varepsilon \leq \frac{1}{1-\gamma}$ , policy  $\widehat{\pi}_p^*$  of perturbed empirical MDP achieves

$$\|V^{\widehat{\pi}_p^*} - V^*\|_\infty \leq \varepsilon \quad \text{and} \quad \|Q^{\widehat{\pi}_p^*} - Q^*\|_\infty \leq \gamma\varepsilon$$

with sample complexity at most

$$\tilde{O}\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\varepsilon^2}\right).$$

- $\widehat{\pi}_p^*$ : obtained by empirical QVI or PI within  $\tilde{O}\left(\frac{1}{1-\gamma}\right)$  iterations
- minimax lower bound:  $\widetilde{\Omega}\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\varepsilon^2}\right)$  [Azar et al., 2013]



**A sketch of the main proof ingredients**

## Notation and Bellman equation

---

- $V^\pi$ : true value function under policy  $\pi$ 
  - ▶ Bellman equation:  $V = (I - \gamma P_\pi)^{-1}r$  [Sutton and Barto, 2018]

## Notation and Bellman equation

---

- $V^\pi$ : true value function under policy  $\pi$ 
  - ▶ Bellman equation:  $V = (I - \gamma P_\pi)^{-1}r$  [Sutton and Barto, 2018]
- $\hat{V}^\pi$ : estimate of value function under policy  $\pi$ 
  - ▶ Bellman equation:  $\hat{V} = (I - \gamma \hat{P}_\pi)^{-1}r$

## Notation and Bellman equation

---

- $V^\pi$ : true value function under policy  $\pi$ 
  - ▶ Bellman equation:  $V = (I - \gamma P_\pi)^{-1}r$  [Sutton and Barto, 2018]
- $\hat{V}^\pi$ : estimate of value function under policy  $\pi$ 
  - ▶ Bellman equation:  $\hat{V} = (I - \gamma \hat{P}_\pi)^{-1}r$
- $\pi^*$ : optimal policy w.r.t. true value function
- $\hat{\pi}^*$ : optimal policy w.r.t. empirical value function

## Notation and Bellman equation

---

- $V^\pi$ : true value function under policy  $\pi$ 
  - ▶ Bellman equation:  $V = (I - \gamma P_\pi)^{-1}r$  [Sutton and Barto, 2018]
- $\hat{V}^\pi$ : estimate of value function under policy  $\pi$ 
  - ▶ Bellman equation:  $\hat{V} = (I - \gamma \hat{P}_\pi)^{-1}r$
- $\pi^*$ : optimal policy w.r.t. true value function
- $\hat{\pi}^*$ : optimal policy w.r.t. empirical value function
- $V^* := V^{\pi^*}$ : optimal values under true models
- $\hat{V}^* := \hat{V}^{\hat{\pi}^*}$ : optimal values under empirical models

## Proof ideas (cont.)

---

Elementary decomposition:

$$V^* - V^{\widehat{\pi}^*} = (V^* - \widehat{V}^{\pi^*}) + (\widehat{V}^{\pi^*} - \widehat{V}^{\widehat{\pi}^*}) + (\widehat{V}^{\widehat{\pi}^*} - V^{\widehat{\pi}^*})$$

## Proof ideas (cont.)

---

Elementary decomposition:

$$\begin{aligned} V^* - V^{\widehat{\pi}^*} &= (V^* - \widehat{V}^{\pi^*}) + (\widehat{V}^{\pi^*} - \widehat{V}^{\widehat{\pi}^*}) + (\widehat{V}^{\widehat{\pi}^*} - V^{\widehat{\pi}^*}) \\ &\leq (V^* - \widehat{V}^{\pi^*}) + \mathbf{0} + (\widehat{V}^{\widehat{\pi}^*} - V^{\widehat{\pi}^*}) \end{aligned}$$

## Proof ideas (cont.)

---

Elementary decomposition:

$$\begin{aligned} V^* - V^{\widehat{\pi}^*} &= (V^* - \widehat{V}^{\pi^*}) + (\widehat{V}^{\pi^*} - \widehat{V}^{\widehat{\pi}^*}) + (\widehat{V}^{\widehat{\pi}^*} - V^{\widehat{\pi}^*}) \\ &\leq (V^* - \widehat{V}^{\pi^*}) + \mathbf{0} + (\widehat{V}^{\widehat{\pi}^*} - V^{\widehat{\pi}^*}) \end{aligned}$$

- **Step 1:** control  $V^\pi - \widehat{V}^\pi$ , for fixed  $\pi$   
**(Bernstein's inequality + high order decomposition)**

## Proof ideas (cont.)

---

Elementary decomposition:

$$\begin{aligned} V^* - V^{\widehat{\pi}^*} &= (V^* - \widehat{V}^{\pi^*}) + (\widehat{V}^{\pi^*} - \widehat{V}^{\widehat{\pi}^*}) + (\widehat{V}^{\widehat{\pi}^*} - V^{\widehat{\pi}^*}) \\ &\leq (V^* - \widehat{V}^{\pi^*}) + \mathbf{0} + (\widehat{V}^{\widehat{\pi}^*} - V^{\widehat{\pi}^*}) \end{aligned}$$

- **Step 1:** control  $V^\pi - \widehat{V}^\pi$ , for fixed  $\pi$   
**(Bernstein's inequality + high order decomposition)**
- **Step 2:** control  $\widehat{V}^{\widehat{\pi}^*} - V^{\widehat{\pi}^*}$   
**(decouple statistical dependence)**

## Step 1: high order decomposition

---

Bellman equation  $V^\pi = (I - \gamma P_\pi)^{-1} r$

## Step 1: high order decomposition

---

Bellman equation  $V^\pi = (I - \gamma P_\pi)^{-1} r$

[Agarwal et al., 2019]  $\widehat{V}^\pi - V^\pi = \gamma(I - \gamma P_\pi)^{-1}(\widehat{P}_\pi - P_\pi)\widehat{V}^\pi$  (\*)

## Step 1: high order decomposition

---

Bellman equation  $V^\pi = (I - \gamma P_\pi)^{-1} r$

[Agarwal et al., 2019]  $\widehat{V}^\pi - V^\pi = \gamma(I - \gamma P_\pi)^{-1}(\widehat{P}_\pi - P_\pi)\widehat{V}^\pi$  (\*)

[ours]  $\widehat{V}^\pi - V^\pi = \gamma(I - \gamma P_\pi)^{-1}(\widehat{P}_\pi - P_\pi)V^\pi +$   
 $+ \gamma(I - \gamma P_\pi)^{-1}(\widehat{P}_\pi - P_\pi) [\widehat{V}^\pi - V^\pi]$

## Step 1: high order decomposition

---

Bellman equation  $V^\pi = (I - \gamma P_\pi)^{-1} r$

[Agarwal et al., 2019]  $\hat{V}^\pi - V^\pi = \gamma(I - \gamma P_\pi)^{-1}(\hat{P}_\pi - P_\pi)\hat{V}^\pi$  (\*)

[ours]  $\hat{V}^\pi - V^\pi = \gamma(I - \gamma P_\pi)^{-1}(\hat{P}_\pi - P_\pi)V^\pi +$   
 $+ \gamma^2((I - \gamma P_\pi)^{-1}(\hat{P}_\pi - P_\pi))^2\hat{V}^\pi$

## Step 1: high order decomposition

---

Bellman equation  $V^\pi = (I - \gamma P_\pi)^{-1} r$

$$[\text{Agarwal et al., 2019}] \quad \widehat{V}^\pi - V^\pi = \gamma(I - \gamma P_\pi)^{-1}(\widehat{P}_\pi - P_\pi)\widehat{V}^\pi \quad (*)$$

$$\begin{aligned} [\text{ours}] \quad \widehat{V}^\pi - V^\pi &= \gamma(I - \gamma P_\pi)^{-1}(\widehat{P}_\pi - P_\pi) \textcolor{red}{V}^\pi + \\ &\quad + \gamma^2 \left( (I - \gamma P_\pi)^{-1}(\widehat{P}_\pi - P_\pi) \right)^2 \textcolor{red}{V}^\pi \\ &\quad + \gamma^2 \left( (I - \gamma P_\pi)^{-1}(\widehat{P}_\pi - P_\pi) \right)^2 [\widehat{V}^\pi - \textcolor{red}{V}^\pi] \end{aligned}$$

## Step 1: high order decomposition

---

Bellman equation  $V^\pi = (I - \gamma P_\pi)^{-1} r$

[Agarwal et al., 2019]  $\hat{V}^\pi - V^\pi = \gamma(I - \gamma P_\pi)^{-1}(\hat{P}_\pi - P_\pi)\hat{V}^\pi \quad (\star)$

[ours]  $\hat{V}^\pi - V^\pi = \gamma(I - \gamma P_\pi)^{-1}(\hat{P}_\pi - P_\pi)V^\pi +$   
 $+ \gamma^2((I - \gamma P_\pi)^{-1}(\hat{P}_\pi - P_\pi))^2 V^\pi$   
 $+ \gamma^3((I - \gamma P_\pi)^{-1}(\hat{P}_\pi - P_\pi))^3 V^\pi$   
 $+ \dots$

## Step 1: high order decomposition

---

Bellman equation  $V^\pi = (I - \gamma P_\pi)^{-1} r$

[Agarwal et al., 2019]  $\hat{V}^\pi - V^\pi = \gamma(I - \gamma P_\pi)^{-1}(\hat{P}_\pi - P_\pi)\hat{V}^\pi \quad (\star)$

[ours]  $\hat{V}^\pi - V^\pi = \gamma(I - \gamma P_\pi)^{-1}(\hat{P}_\pi - P_\pi)V^\pi +$   
 $+ \gamma^2((I - \gamma P_\pi)^{-1}(\hat{P}_\pi - P_\pi))^2 V^\pi$   
 $+ \gamma^3((I - \gamma P_\pi)^{-1}(\hat{P}_\pi - P_\pi))^3 V^\pi$   
 $+ \dots$

Bernstein's inequality:  $|(\hat{P}_\pi - P_\pi)V^\pi| \leq \sqrt{\frac{\text{Var}[V^\pi]}{N}} + \frac{\|V^\pi\|_\infty}{N}$

## Byproduct: policy evaluation

Theorem (Li, Wei, Chi, Gu, Chen'20)

Fix any policy  $\pi$ . For every  $0 < \varepsilon \leq \frac{1}{1-\gamma}$ , plug-in estimator  $\hat{V}^\pi$  obeys

$$\|\hat{V}^\pi - V^\pi\|_\infty \leq \varepsilon$$

with sample complexity at most

$$\tilde{O}\left(\frac{|\mathcal{S}|}{(1-\gamma)^3 \varepsilon^2}\right).$$

## Byproduct: policy evaluation

Theorem (Li, Wei, Chi, Gu, Chen'20)

Fix any policy  $\pi$ . For every  $0 < \varepsilon \leq \frac{1}{1-\gamma}$ , plug-in estimator  $\widehat{V}^\pi$  obeys

$$\|\widehat{V}^\pi - V^\pi\|_\infty \leq \varepsilon$$

with sample complexity at most

$$\widetilde{O}\left(\frac{|\mathcal{S}|}{(1-\gamma)^3\varepsilon^2}\right).$$

- minimax lower bound [Azar et al., 2013, Pananjady and Wainwright, 2019]

## Byproduct: policy evaluation

Theorem (Li, Wei, Chi, Gu, Chen'20)

Fix any policy  $\pi$ . For every  $0 < \varepsilon \leq \frac{1}{1-\gamma}$ , plug-in estimator  $\widehat{V}^\pi$  obeys

$$\|\widehat{V}^\pi - V^\pi\|_\infty \leq \varepsilon$$

with sample complexity at most

$$\widetilde{O}\left(\frac{|\mathcal{S}|}{(1-\gamma)^3\varepsilon^2}\right).$$

- minimax lower bound [Azar et al., 2013, Pananjady and Wainwright, 2019]
- tackle sample size barrier: prior work requires sample size  $> \frac{|\mathcal{S}|}{(1-\gamma)^2}$   
[Agarwal et al., 2019, Pananjady and Wainwright, 2019, Khamaru et al., 2020]

## Step 2: controlling $\hat{V}^{\hat{\pi}^*} - V^{\hat{\pi}^*}$

---

A natural idea: apply our policy evaluation theory + union bound

## Step 2: controlling $\hat{V}^{\hat{\pi}^*} - V^{\hat{\pi}^*}$

---

A natural idea: apply our policy evaluation theory + union bound

- highly suboptimal!

## Step 2: controlling $\hat{V}^{\hat{\pi}^*} - V^{\hat{\pi}^*}$

---

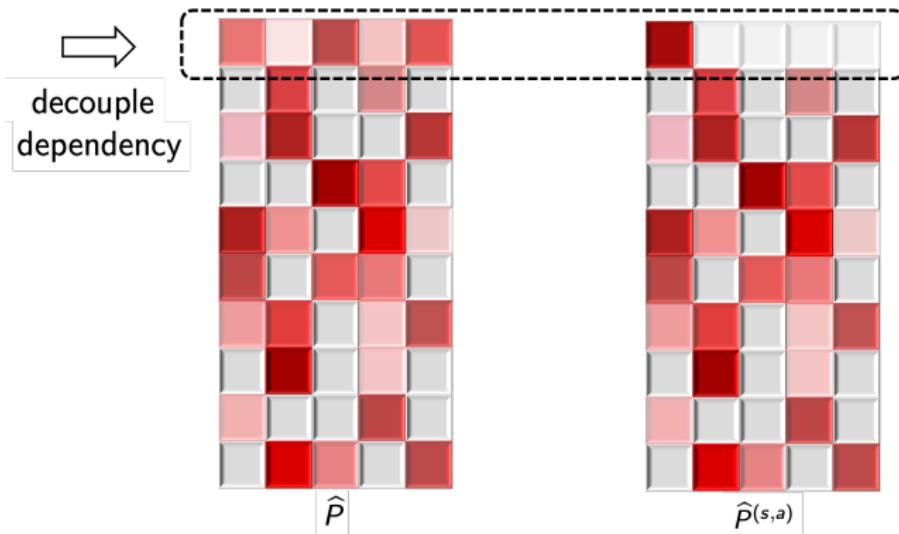
A natural idea: apply our policy evaluation theory + union bound

- highly suboptimal!

**key idea 2:** a **leave-one-out argument** to decouple stat. dependency btw  $\hat{\pi}$  and samples

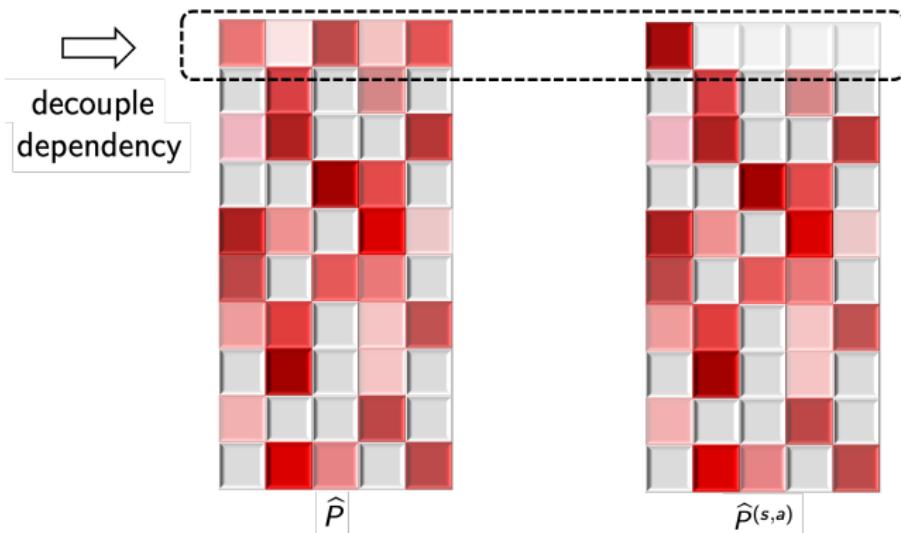
— *inspired by [Agarwal et al., 2019] but quite different ...*

## Key idea 2: leave-one-out argument



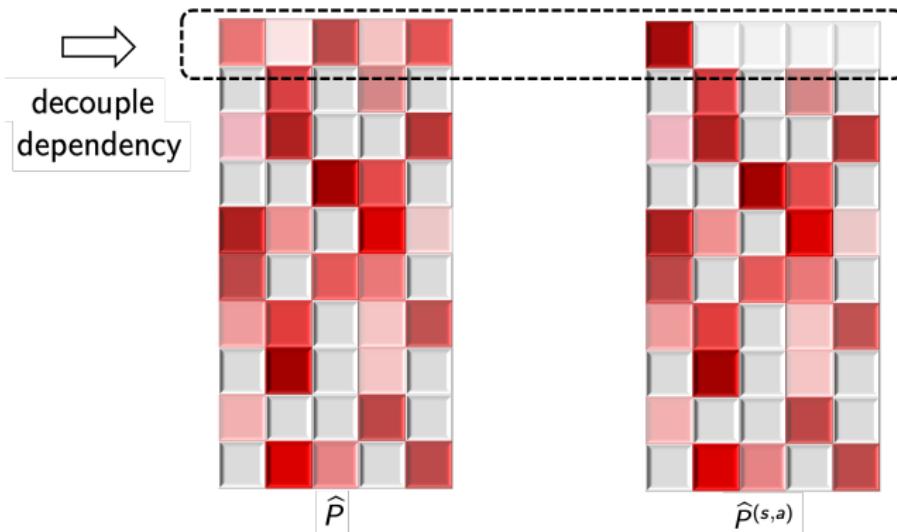
- state-action absorbing MDP for each  $(s, a)$ :  $(\mathcal{S}, \mathcal{A}, \hat{P}^{(s,a)}, r, \gamma)$

## Key idea 2: leave-one-out argument



- state-action absorbing MDP for each  $(s, a)$ :  $(\mathcal{S}, \mathcal{A}, \hat{P}^{(s,a)}, r, \gamma)$
- $(\hat{P} - P)_{s,a} V_{\hat{\pi}^*} = (\hat{P} - P)_{s,a} V_{\hat{\pi}_{s,a}^*}$  ( $\hat{\pi}_{s,a}^*$ : optimal for new MDP)

## Key idea 2: leave-one-out argument



- state-action absorbing MDP for each  $(s, a)$ :  $(\mathcal{S}, \mathcal{A}, \hat{P}^{(s,a)}, r, \gamma)$
- $(\hat{P} - P)_{s,a} V_{\hat{\pi}^*} = (\hat{P} - P)_{s,a} V_{\hat{\pi}_{s,a}^*}$  ( $\hat{\pi}_{s,a}^*$ : optimal for new MDP)

**Caveat:** require  $\hat{\pi}^*$  to stand out from other policies

## Key idea 3: tie-breaking via perturbation

---

- How to ensure the optimal policy stand out from other policies?

$$\forall s \in \mathcal{S}, \quad \widehat{Q}^*(s, \widehat{\pi}^*(s)) - \max_{a: a \neq \widehat{\pi}^*(s)} \widehat{Q}^*(s, a) \geq \omega$$

## Key idea 3: tie-breaking via perturbation

---

- How to ensure the optimal policy stand out from other policies?

$$\forall s \in \mathcal{S}, \quad \widehat{Q}^*(s, \widehat{\pi}^*(s)) - \max_{a: a \neq \widehat{\pi}^*(s)} \widehat{Q}^*(s, a) \geq \omega$$

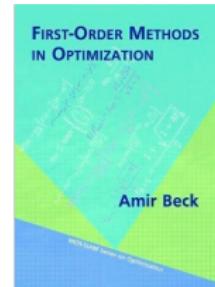
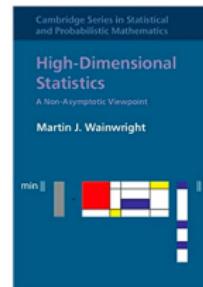
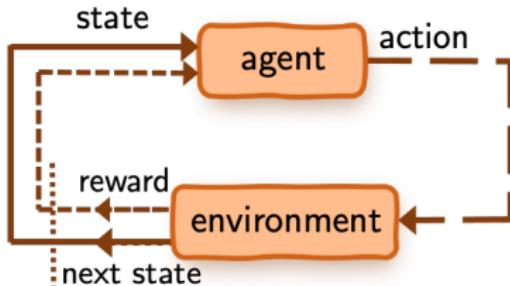
- **Solution:** slightly perturb rewards  $r \implies \widehat{\pi}_p^*$

- ▶ ensures the uniqueness of  $\widehat{\pi}_p^*$
- ▶  $V^{\widehat{\pi}_p^*} \approx V^{\widehat{\pi}^*}$



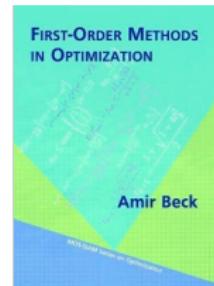
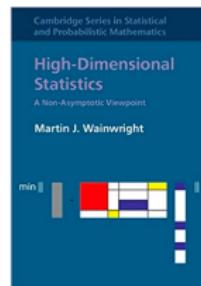
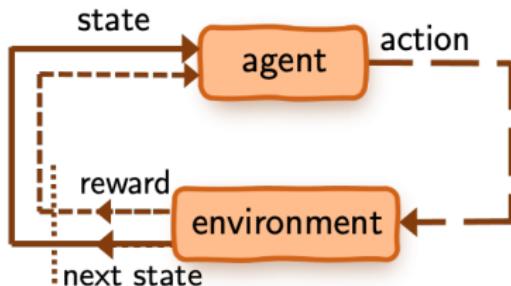
# Concluding remarks

Understanding RL requires modern statistics and optimization



# Concluding remarks

Understanding RL requires modern statistics and optimization



## Future directions

- beyond the tabular setting  
[Feng et al., 2020, Jin et al., 2019, Duan and Wang, 2020]
- finite-horizon episodic MDPs  
[Dann and Brunskill, 2015, Jiang and Agarwal, 2018, Wang et al., 2020]

## **Paper:**

“Breaking the sample size barrier in model-based reinforcement learning with a generative model,” G. Li, Y. Wei, Y. Chi, Y. Gu, Y. Chen, arxiv:2005.12900, 2020